Statistical Learning and Data Analytics, Spring 2017
Take-home Final Exam

(Due on June 19th, 2017)

Instructions:

i. *This exam is take-home exam. You MUST finish the exam INDEPENDENTLY without discussing with others.*

ii. *Show all your work to justify your answers. Answers without adequate justification will not receive credit.*

iii. *For the computational problems, try to make inferences on the computer outputs when answering the questions. Do not only paste the computer outputs.*

iv. *Email your answer sheets to me (xdeng@vt.edu) by* ***10:00am on June 19, 2017****, togherther with your R codes.*

Problem 1

Let $X_1 \in \mathbb{R}$ and $X_2 \in \mathbb{R}$ be random variables and

$$Y = m(X_1, X_2) + \varepsilon$$

where $E(\varepsilon) = 0$ and $E(\varepsilon^2) = \sigma^2$. Consider the class of multiplicative predictors of the form $m(x_1, x_2) = \beta x_1 x_2$. Let β^* be the best predictor, that is, β^* minimizes $E_{Y,X_1,X_2}(Y - \beta X_1 X_2)^2$. Find an expression for β^*.

Problem 2

Suppose that $x_1, \ldots, x_n$ are an independent and identically distributed (i.i.d.) sample from a Bernouli distribution with parameter p as follows,

$$X = \begin{cases} 1, & \text{with probablility } p \\ -1, & \text{with probability } 1-p \end{cases}$$

Please derive the maximum likelihood estimator of p.

Y' = (y+1)/2

Computational problems

Problem 3 Ridge and Lasso regression
Use the LA ozone dataset. Divide the dataset into two groups at random. One group, which we call the training data, containing 2/3 of the observations and one group, which we call the test data, with 1/3 of the observations. In the following you are asked to regress the *cube root* of the ozone on the other variables. You should *only* use the training data for the estimation.

(a) Best subset model: find the best subset model for each model size, i.e., the number of variables included, $p = 1, 2, \ldots, 9$, according to the C_p criterion. Return two plots: (1) C_p value with respect to the degree of freedom $p+1$, $p = 1, 2, \ldots, 9$ (2) Training error=$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ with respect to $p+1$ and Test error=$\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$ for the test data set w.r.t. $p+1$. Here $p = 0, 1, \ldots, 9$.

(b) Lasso method: use the Lasso method and plot the Training error and Test error with respect to λ vary from small to large.

(c) Ridge regression: use ridge regression and plot the Training error and Test error with respect to λ vary from small to large.

Problem 4 Classification
The traing and test data sets can be found from 'geno_train.txt' and 'geno_test.txt'. Each contain 16 columns of data from different individuals, with the first 15 being the genetic fingerprint (the count of the number of repeats for certain so-called tandem repeats in the genome) and the last being the population variable. The purpose is to predict the population from the genetic fingerprint. We refer below to the repeat counts as the count data (the x variables) and the population as the group (the y variable).

(a) Use the logisitic regression to analyze the data, and calculate the misclassification rate on the test data set.

(b) Use the LDA method to analyze the data, and calculate the misclassification rate on the test data set.